# <Bioinformatics Center>Chemical Lfe Science

# Bioinformatics Center
## – Chemical Life Science –

http://cls.kuicr.kyoto-u.ac.jp/

Assoc Prof
GOTO, Susumu
(D Eng)

Program-Specific Assist Prof
TOKIMATSU, Toshiaki
(D Agr)

Program-Specific Assist Prof
KOTERA, Masaaki
(D Sc)

PD
JOANNIN, Nicolas
(Ph D)

### Researchers

MORIYA, Yuki
NAKAGAWA, Zenichi
MUTO, Ai

### Students

SHIMIZU, Yugo (D3)
MIZUTANI, Sayaka (D3)
NISHIMURA, Yosuke (D3)
FUJITA, Megumi (D3)
MIHARA, Tomoko (D3)
JIN, Zhao (D3)
SATOH, Takanori (M2)

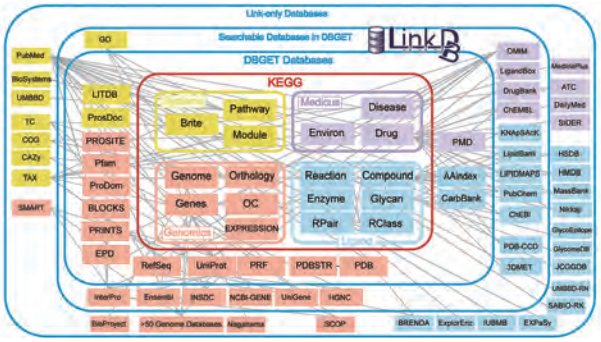
### Visiting Researcher

Assoc Prof INGALE, Arun Govindrao North Maharashtra University, India, 1 May–31 July

## Scope of Research

The proteins responsible for biosynthesis, biodegradation, and transport of additional molecules, such as small metabolites, lipids and glycans, are encoded in the genome, which may indicate that all cellular functions are specified by the genomic DNA sequence. In practice, however, inferring higher-level systemic functions of the cell or the organism needs more than solely the genomic information. We are developing bioinformatics methods to integrate different types of data and knowledge on various aspects of the biological systems towards basic understanding of life as a molecular interaction/reaction system and also towards practical applications in medical and pharmaceutical sciences.


GenomeNet Top page




Databases available in the DBGET/LinkDB system of the GenomeNet service. Color of each database represents the type of its contents, yellow: systems information, purple: medical information, pink: genetic information, light blue: chemical information.

### KEYWORDS

GenomeNet
(Meta)genomics
Reaction Ontology
Bioinformatics
Pharmacoinformatics

### Selected Publications


Kotera, M.; Tabei, Y.; Yamanishi, Y.; Tokimatsu, T.; Goto, S., Supervised *de novo* Reconstruction of Metabolic Pathways from Metabolome-scale Compound Sets, *Bioinformatics*, **29,** i135-i144 (2013).

Muto, A.; Kotera, M.; Tokimatsu, T.; Nakagawa, Z.; Goto, S.; Kanehisa, M., Modular Architecture of Metabolic Pathways Revealed by Conserved Sequences of Reactions, *J. Chem. Inf. Model*, **53,** 613-622 (2013).

Nakaya, A.; Katayama, T.; Itoh, M.; Hiranuka, K.; Kawashima, S.; Moriya, Y.; Okuda, S.; Tanaka, M.; Tokimatsu, T.; Yamanishi, Y.; Yoshizawa, A. C.; Kanehisa, M.; Goto, S., KEGG OC: A Large-scale Automatic Construction of Taxonomy-based Ortholog Clusters, *Nucleic Acids Res.*, **41,** D353-D357 (2013).

Mizutani, S.; Pauwels, E.; Stoven, V.; Goto, S.; Yamanishi, Y., Relating Drug-protein Interaction Network with Drug Side-effects, *Bioinformatics*, **28,** i522-i528 (2012).

Kotera, M.; Yamanishi, Y.; Moriya, Y.; Kanehisa, M.; Goto, S., GENIES: Gene Network Inference Engine Based on Supervised Analysis, *Nucleic Acids Res.*, **40,** W162-W167 (2012).

## KEGG OC: An Automatically Constructed Comprehensive Ortholog Clusters

As the number of fully sequenced genomes is rapidly growing, it has become increasingly important to automate the identification of orthologs and the construction of ortholog clusters in order to understand functional properties and biological roles of genes in these genomes. The ortholog clusters also play a key role in functional annotation for newly sequenced genomes, because orthologs tend to have equivalent functions. KEGG OC is a novel database of ortholog clusters based on the whole genome comparison. The current version of KEGG OC contains 1,223,674 ortholog clusters, obtained by clustering 10,615,995 genes in 2,717 complete genomes. The ortholog clusters in KEGG OC were constructed by applying a novel clustering method to all possible protein coding genes in all complete genomes, based on their amino acid sequence similarities. The originality of our clustering algorithm lies in the use of a quasi-clique search (Figure 1) and a step-by-step clustering along the phylogenetic tree. In comparison with KEGG ORTHOLOGY (KO), which is not comprehensive but curated manually, almost every ortholog cluster contains a single major KO group. It suggests that the KEGG OC contains high-quality ortholog clusters. Specifically, KEGG OC has the following features. First, it consists of all fully sequenced genomes of a wide range of organisms from eukaryotes, bacteria and archaea. Second, the ortholog clusters have a hierarchical structure based on the step-by-step clustering. Third, it is computationally efficient to calculate the comprehensive ortholog clusters, which makes it possible to regularly update the contents. Forth, it is compatible with the KEGG database, which provides an easy way to link the ortholog clusters with KEGG PATHWAY, BRITE functional hierarchies, and many more.

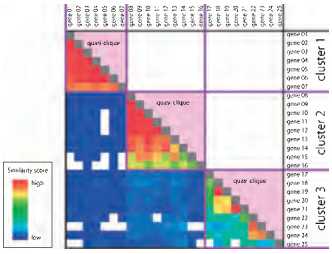



**Figure 1.** Overview of quasi-clique search on the adjacency matrix.

## Modular Architecture of Metabolic Pathways Revealed by Conserved Sequences of Reactions

With increasing complexity of metabolic networks stored in databases, the need to describe the functional modules of metabolic networks is growing. Modular descriptions of network help us to understand complex network by split them into functionally related sub-networks. Towards better understanding of the evolution of the metabolic network, we developed a method to extract conserved sequences of reactions called 'reaction modules' from the analysis of all known metabolic pathways stored in the KEGG PATHWAY database. Our method is based on the similarity of chemical structure transformation patterns along the metabolic pathways. This is a purely chemical similarity measure without incorporating any protein sequence information or the EC number information. It enables us to analyze reactions with no EC numbers assigned or even with no enzymes identified. The reaction modules, which are conserved sequences of similar reactions, are systematically searched in the KEGG metabolic pathways using the similarity scoring scheme between reactions. The extracted reaction modules are repeatedly used as if they are building blocks of the metabolic network and contain chemical logic of organic reactions. We identified well-conserved, possibly ancient, reaction modules involving 2-oxocarboxylic acids (Figure 2). The chain extension module that appears as the tricarboxylic acid reaction sequence in the TCA cycle is now shown to be used in other pathways together with different types of modification modules. We also identified reaction modules and their connection patterns for aromatic ring cleavages in microbial biodegradation pathways, which are most characteristic in terms of both distinct reaction sequences and distinct gene clusters. The modular architecture of biodegradation modules will have a potential for predicting degradation pathways of xenobiotic compounds. The collection of these and many other reaction modules is made available as part of the KEGG database.

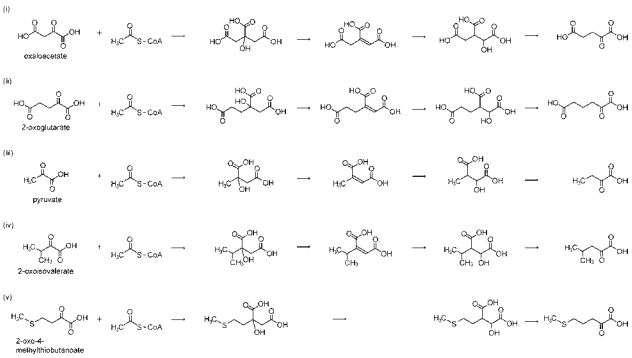



**Figure 2.** 2-oxocarboxylic acid chain extension module. Examples of reaction module RM001, which performs chain extension by consuming one acetyl-CoA. (i) Extension from oxaloacetate to 2-oxoglutarate in citrate cycle. (ii) 2-oxoglutarate to 2-oxoadipate in lysine biosynthesis. (iii) pyruvate (2-oxopropanoate) to 2-oxobutanoate and (iv) 2-oxoisovalerate to 2-oxoisocaproate in valine, leucine and isoleucine biosynthesis. (v) 2-oxo-4-methylthiobutanoate to 2-oxo-5-methylthiopentanoic acid, the first reaction module of a six tandem repeat of RM001 toward 2-oxo-10-methylthiodecanoate in the biosynthesis pathway of glucosinolates.